# Ridesourcing Behavior Profiles: Application of K-Prototype Analysis on Large-scale Data from Chicago, Illinois

**Jason Soria**
PhD Candidate
Northwestern University
Department of Civil and Environmental Engineering
Technological Institute, 2145 Sheridan Road, Evanston, IL 60208 USA
jason.soria@u.northwestern.edu

**Ying Chen**
Research Assistant Professor
Northwestern University
Department of Civil and Environmental Engineering
Transportation Center, 600 Foster St., Evanston, IL 60208 USA
y-chen@northwestern.edu

**Amanda Stathopoulos**
Assistant Professor
Northwestern University
Department of Civil and Environmental Engineering
Technological Institute, 2145 Sheridan Road, Evanston, IL 60208 USA
a-stathopoulos@northwestern.edu

Word Count: 197 (abstract) + 4394 (text) + 869 (references) + 250*(2 tables) = 5960

Submitted for presentation and publication to ADB40 (Standing Committee on Transportation Demand Forecasting) and ABE30 (Standing Committee on Transportation Issues in Major Cities) including the ABE30 call for papers titled "Designing Cities for New Mobility Options" at the 99th Annual Meeting of the Transportation Research Board, January 2020.

**August 1, 2019**

**ABSTRACT**

Shared mobility-on-demand services are evolving rapidly in cities around the world. As a prominent example, ridesourcing is becoming an integral part of many urban transportation ecosystems. Despite the centrality, limited public availability of detailed temporal and spatial data on ridesourcing trips has stifled research in how new services interact with traditional mobility options and how they impact travel in cities. Improving data-sharing agreements is opening unprecedented opportunities for research in this area. This study's goal is to study emerging patterns of mobility using the recently released City of Chicago public ridesourcing data. The data are supplemented with weather, transit, and taxi data to gain a broader understanding of ridesourcing's role in the mobility ecosystem. Considering the analysis data is large and contains numerical and categorical variables, K-prototypes is utilized for its ability to accept mixed variable type data. An extension of the K-means algorithm, its output is a classification of the data into several clusters called prototypes. Six ridesourcing prototypes were identified, described, and discussed in this study. Identified user segments are defined by adverse weather conditions, competition with alternative modes, spatial patterns, and tendency for ridesplitting.

*Keywords:* Transportation network company, Ridesourcing, Ridesplitting, K-prototype, Clustering analysis

## INTRODUCTION

Transportation Network Companies (TNCs) are prominent in many urban transportation ecosystems. The growth of ridesourcing services is attributed to improvements in Information and Communication Technologies that made these services more convenient to use compared to traditional modes of transport. However, the proliferation of ridesourcing services has been a disruptive force as it transforms the mobility landscape. This transformation has not been widely studied as many TNCs are reluctant to make their data publicly available. Recent data-sharing agreements with the City of Chicago, IL enables researchers to examine the role of ridesourcing in the transportation ecosystem using an abundance of temporal and spatial data on ridesourcing trips.

Several studies have characterized the adoption, frequency, and attitudes towards ridesourcing (Circella et al., 2016, Dias et al., 2017, Alemi et al., 2018a, Alemi et al., 2018b), but none have used publicly available trip data at the scale and scope provided by the City of Chicago. This study uses this newly released trip data, to develop insights about the role ridesourcing plays in the transportation ecosystem. The detailed data from operators Uber, Lyft and Via is fused with local transit, and taxi data, as well as weather observations. The purpose of this research is to study the mobility patterns present in the data by grouping similar trips together.

This paper utilizes an unsupervised learning algorithm to examine the underlying relationships in the data. Due to the mixed data types (i.e. data containing both numeric and qualitative/categorical variables), a clustering algorithm must be chosen carefully. The unsupervised learning technique proposed in this paper is the K-Prototypes algorithm developed by Huang (1998). It is an extension of the K-Means algorithm that accepts categorical data. The results of this model are similar to K-Means algorithm as the output is a classification of the data into K number of prototypes, the equivalent of clusters. A more complete explanation of model development and attribute selection is given in the methodology section.

This study contributes to the literature by providing a closer look at large, relatively disaggregate TNC data in a major metropolitan area. After tuning parameters for the best fit, the optimal number of prototypes to describe the data is 6. The first group of users (i.e. prototype) contains trips that occur in adverse weather conditions such as rainy weather. The second prototype involves trips that occur in the evening. The third prototype represents trips that are typically longer in distance but are not shared. The fourth prototype is defined by trip origin and destinations being to the two major airports in Chicago: O'Hare and Midway. The fifth prototype is defined by short, not shared trips occurring in areas that are well served by transit. The sixth prototype is defined by nearly all observations being shared rides.

This study is structured with the following sections. Following this section is a literature review that covers the state-of-the-art in TNC research. After the literature review is the methodology section which gives an overview of the K-Prototypes algorithm and how it fits the purpose of this study. The next section reviews the algorithm's output and leads into the discussion. Finally, the conclusion contains a review of what is achieved in this study, its limitations, and possible future works.

## LITERATURE REVIEW

TNCs as they are known now were introduced with the inception of Uber in 2009 and Lyft shortly after. Ridesharing has existed long before these companies came about, but the innovations that they brought via improved communications technology to allow immediate street-hailing has radically altered the transportation ecosystem. There are several levels of TNC services that are

defined by Shaheen and Cohen (2018). TNC provided trips are considered ridesourcing trips and should not be confused with ridesharing trips. Unlike the ridesharing case where drivers participate to offset trip costs, TNC drivers are servicing the customer request in exchange for a fare. More recently, new trip categories have emerged within ridesourcing, such as splitting rides or curb-to-curb travel. For the major TNCs, riders can decide in most cases to share their trip with other parties that are traveling along the same trajectory. Authorizing this typically results in lower fares but longer travel times as the trip now includes several stops that may cause the vehicle to deviate from the optimal path for a single origin-destination pair. This phenomenon is called ridesplitting. More definitions of TNC provided trips can be found in Shaheen and Cohen (2018).

Researchers have tried to develop a better understanding of ridesourcing trips, but data is scarce. Uber and Lyft do not publicly share their data so there has been a dearth of empirical studies. Because the data is limited, Henao and Marshall (2018) went so far as to become a TNC driver and collect trip information themselves. They found that ridesourcing is not efficient. After accounting for deadheading mileage, the average occupancy of a TNC vehicle is less than one person. The extra vehicle miles traveled caused by deadheading is also a prominent result from their analysis.

The current understanding of ridesourcing travel is mostly informed by survey research. In the following we briefly overview relevant ridesourcing work and relate findings to the current analysis of real large-scale data. Several studies delve into the trip purposes of ridesourcing trips. Defined by its utilization of large capacity vehicles, micro-transit (also known as demand-responsive transit, on-demand transit, or flexible transit) can serve as a tool to address public transit overcrowding and the first-last mile problem (Shaheen and Chan, 2016). It is mostly utilized to commute (Shaheen and Chan, 2016, Lewis and MacKenzie, 2017). Trips made by the more taxi-like TNCs are mostly for social/recreational trips (Rayle et al., 2014, Zhen, 2015, Mahmoudifard et al., 2017, Henao and Marshall, 2018). Trip purpose is not included in the current analysis due to the data anonymization. However, in future works spatial examination of locations of interest combined with other trip attributes can be used to infer trip types.

The effects of TNCs on the transportation system is a core area of research. In particular, due to the similarity of the services, the impact on taxis has been widely studied. TNCs have significantly reduced the demand for traditional taxi services such that taxi drivers altered their strategies to remain profitable (Nie, 2017, Kim et al., 2018, Contreras and Paz, 2018, Jiang and Zhang, 2018, Dong et al., 2018, Berger et al., 2018). Schwieterman and Smith (2018) also find that TNCs are preferred over public transit especially when origin-destination pairs are not well served by transit. Further determinants of ridesourcing use relate to the travel environment. Frei et al. (2017) found that weather affects TNC usage. Though their study focused on micro-transit, TNC services may also be affected by adverse weather.

Owing to these observations, the ridesourcing trip data used in this project are supplemented by data on weather, equivalent transit travel times, and peak taxi demand.

## MODEL DEVELOPMENT

The data analysis used to examine patterns of ridesourcing use in this project is an unsupervised learning technique called K-Prototypes. K-prototypes is similar to K-means since both aim to cluster several observations together according to their attributes. The advantage K-prototypes has in this situation is its ability to also accept categorical variables. More details on K-prototypes development can be found in Huang (1998).

The challenge of dealing with categorical variables has been considered for segmentation analysis. The problem is that the K-means algorithm relies on all variables to be numerical. Specifically, in the K-means algorithm for a continuous variable such as travel time, the distance between an observation's travel time and the proposed cluster's mean travel time is the key element for identifying clusters among observations. With a categorical variable such as vehicle type, the distance is no longer applicable. One strategy to include categorical variables in the K-means algorithm is to code each category as a dummy variable (0 or 1). The distance calculated by K-means algorithm for a categorical variable is then 0 or 1 because it was coded as a dummy variable, but this no longer makes sense. With the K-prototypes algorithm the mode of the category is used and a measure of a simple matching coefficient is used. The formulation from Huang (1998) of K-prototypes algorithm is summarized in equations 1 to 4.

The matching of observations to prototypes involves reducing the error or cost function. This cost function represents the distance between observation data and the assigned prototype center. **Equation 1** shows that the error, *E*, is the sum of distances from the prototype center. $X_i$ are the attributes of trip *i*, $Q_l$ is the center of prototype l, and $y_{il}$ is a dummy variable that is equal to 0 when trip *i* is assigned to prototype l. It is then the sum of squared distances for n TNC trips across k number of prototypes. **Equation 2** breaks down $d(X_i, Q_l)$ into numerical and categorical components, where the first term is the squared numerical distance of attribute *j* of trip *i* from the center for attribute *j* of prototype l; the second term includes a term to determine the weight, $\gamma_l$, of the categorical variables to the total error *E.* The error of prototype l is then calculated in **Equation 3**, where $E_l^c$ is further explained by **Equation 4**. $C_j$ is the set of all unique values of categorical attribute $j$, and $p(c_j \in C_j \,|l)$ is then the probability of unique value $q_j$ from set $C_j$ being in prototype $l$.

$$E = \sum_{l=1}^{k}\sum_{i=1}^{n} y_{il} d(X_i, Q_l) \tag{1}$$

$$d(X_i, Q_l) = \sum_{j=1}^{m_r} \left(x_{ij}^r - q_{lj}^r\right)^2 + \gamma_l \sum_{j=1}^{m_c} \delta\left(x_{ij}^c, q_{lj}^c\right) \tag{2}$$

$$E_l = \sum_{i=1}^{n} y_{il} \sum_{j=1}^{m_r} \left(x_{ij}^r - q_{lj}^r\right)^2 + \gamma_l \sum_{i=1}^{n} y_{il} \sum_{j=1}^{m_c} \delta\left(x_{ij}^c, q_{lj}^c\right) = E_l^r + E_l^c \tag{3}$$

$$E_l^c = \gamma_l \sum_{j=1}^{m_c} n_l \left(1 - p\left(q_{lj}^c \in C_j \middle| l\right)\right) \tag{4}$$

The advantage of using K-Prototypes algorithm over other clustering algorithms is highlighted by **Equation 4**. A common way to code categorical variables for other data-driven methods is to use one-hot encoding. Using this method, the unique values of a category are coded as a dummy variable where they are equal to 1 when denoting the variable of interest and 0 otherwise. Algorithms using one-hot encoded data fail to recognize that these unique values belong to a categorical variable because categories are reduced to 0 or 1. The advantage of K-Prototype algorithm is then its recognition of these values being part of one categorical variable and using the probability of a unique value from a set $C_j$ being in prototype $l$.

This model is implemented and tuned with the R programming language using the 'clustMixType' package (R Development Core Team, 2008, Szepannek and Aschenbruck, 2019). Using this package, the error is minimized and the weighting of the categorical error is optimized. Much like other clustering methods, the number of prototypes is a tunable parameter. The final tunable parameters are discussed in the results section.

## DATA DESCRIPTION

The data used in this project is a partition of the entire available TNC trip data provided by the City of Chicago (City of Chicago, 2019). The trip data begins on November 1, 2018 and is updated monthly. For the purpose of obtaining lower optimization times and being able to match the equivalent transit travel times, the data is partitioned to weekdays in November 2018. Holidays are not included. This leaves a total of 3,085,070 trips in the dataset. The trips are grouped at the census tract level and include variables such as travel time, travel distance, fare, whether it was a shared trip (and if it was how many other trips were included), census tract origin-destination pairs, and timestamp of pickup and drop-off rounded to the nearest 15-minute increment.

The weather data were collected from OpenWeatherMap specifically for the City of Chicago in November 2018 (Open Weather Map, 2019). The data is at the hourly level and includes amount of rain and snow in the previous hour, qualitative description of the weather (such as raining, hazy, sunny, etc.), and temperature. The station collecting the data is located at O'Hare International Airport at the northwest tip of city limits. The supplementary transit travel times dataset was created for each unique origin-destination-time-day tuple. Transit travel time estimates were obtained using the Google Distance Matrix (Advanced) API by providing the census tract of origin, the census tract of destination, travel mode (transit), and departure time (Google, n.d). From the API, approximate transit travel times between origin-destination pairs are collected. Since the data were only collected from 6AM to 10PM, the TNC trips data are also restricted to these hours. The second piece of supplementary data are the monthly taxi trips between census tract origin-destination pairs. The data used are when taxi demand is at its highest point in 2014 (Chen et al., 2018). This data is also collected and made publicly available by the City of Chicago. These data are further described by Chen et al. (2018).

## ANALYSIS OF RESULTS

The K-prototype algorithm was tuned to select the optimal number of prototypes. This was determined by developing models with number of prototypes ranging from 2 to 14 and calculating the total cost across all observations. The final number of prototypes chosen is 6 based on interpretability of segmentation variables and guidance from the plot which in figure 1 shows a clear "elbow" where there are 6 prototypes. This "elbow" method is a tool for researchers to find an appropriate number of clusters (Madhulatha, 2012). An elbow occurs when adding more clusters does not sufficiently improve the objective function. $\gamma$ is the tradeoff between numerical cost and categorical cost was optimized by the 'kproto' function in the 'clustMixType' package and is estimated to be 1.33 for all prototypes as per **Equation 2** and **4** (Szepannek and Aschenbruck, 2019). There is no intuitive meaning to this value except that it can be user-specified, and higher values mean that the categorical variables are weighted more. **Figure 1** shows how many observations belong in each prototype cluster. The clustering results are shown in **Table 1** along with mean values of explanatory attributes in each prototype. A summary of the top 6 origin and destinations, respectively, are given in **Table 2**.

The analysis did not produce prototypes that are heavily differentiated by temperature or snow fall in the past hour. The first segment of users (Prototype 1 or P1) is the second largest and is characterized by its relatively short travel time, low distance, and low total fare charges. This short-distance travel is coupled with the strongest weather impacts observed, namely the presence of adverse weather seen with rain, humidity, and wind speed. The distinct nature of prototype 1 suggests the use of ridesourcing for short distance travel to cope with adverse weather in the early part of the day. Prototype 2 (P2) also has lower-than-average travel times, distances, and total fares, but is distinct from prototype 1 due to the trip timing in the evening and the lack of relationship to weather conditions. Observing Table 2, these trips are most heavily focused in the wealthy downtown and near north areas. Prototype 3 (P3) has longer travel times which tend to be associated with longer distances (albeit not associated with airport travel) and higher total charges. This large user segment suggests some transit gap-filling capacity of ridesourcing in Chicago whereas the potentially available transit trip would take 30% longer on average with transit travel-time taken as base. Notably, considering the fixed transit pricing, the ridesourcing trips were on average six times more costly. Trips in this prototype are also typically not shared. Prototype 4 (P4) represents a small group of users that also have long travel times but examining **Table 2** shows that the main origin and destinations of these trips is to O'Hare or Midway International Airports. This prototype also has trips where the origins and destinations are not served well by transit as seen with the average transit travel time being more than 40% longer. Along with poor transit connectivity, this cluster features relatively lower taxi frequency. These trips' fares are more expensive than in other prototypes, but relative to the cost of traditional taxis, it is still cheaper. With low taxi frequency, this prototype highlights an advantage of TNCs over taxi. Interestingly, prototype 5 (P5) is a small cluster that stands out as representing the shortest trips and for being the only case where trips could have been served better by transit. Notably, the average transit travel times would be 14% lower than the observed TNC travel times. Most of these trips are in the Chicago Loop or just north of it. Prototype 6 (P6) is defined by representing nearly all shared authorized trips. This segment appears to reflect a more cost-conscious user group given that the ridesourcing price per mile is the lowest, and the competition in terms of price and time is closer to the potentially available transit trip.

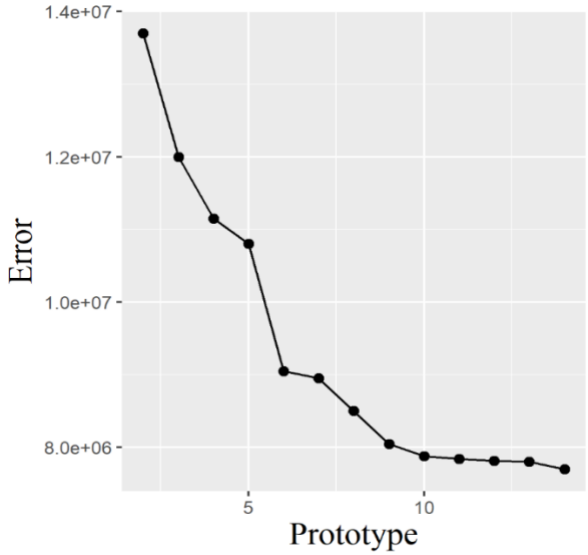


**FIGURE 1 Selection K number of Prototypes**

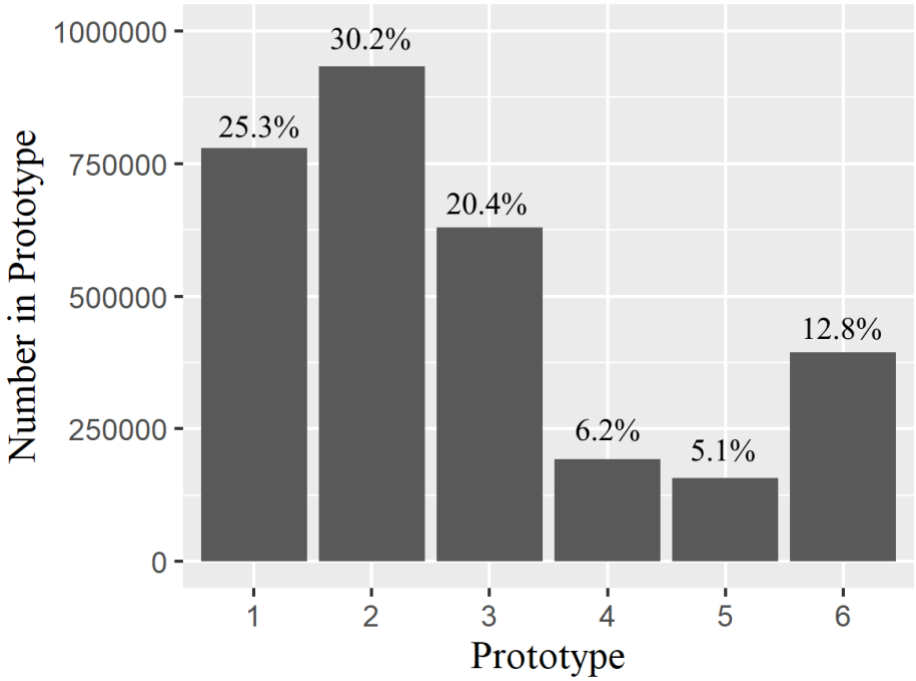


**FIGURE 2 Prototype Shares among total ridesourcing trips**

**TABLE 1 Prototype Attribute results**

| Variable | P1 | P2 | P3 | P4 | P5 | P6 |
|---|---|---|---|---|---|---|
| Travel Time (seconds) | **637.40*** | **600.9** | **1284.0** | **2014.0** | **572.2** | 1320.0 |
| Distance (miles) | **2.16** | **2.07** | **5.74** | **12.25** | **1.39** | 4.83 |
| Total Fare ($) | 9.05 | 8.85 | **15.56** | **27.64** | 8.40 | 7.43 |
| Parties Joined in Trip | 1.07 | 1.07 | 1.04 | 1.17 | 1.12 | **3.00** |
| Humidity (%) | **82.96** | 66.07 | 72.29 | 75.09 | 73.64 | 74.12 |
| Wind Speed (mph) | **4.32** | 3.57 | 3.66 | 3.96 | 3.70 | 3.66 |
| Rain last hour (inches) | **0.15** | 0.01 | 0.03 | 0.08 | 0.05 | 0.05 |
| Minute after Midnight | 702.5 | **1080.0** | 878.0 | 826.4 | 815.9 | 870.7 |
| Transit Travel Time (sec) | 844 | 804 | **1838** | **3392** | **501** | **1545** |
| Monthly Taxi Frequency | 11695 | 10031 | **3558** | **3840** | **144687** | **5286** |
| Percent ridesplitting (%) | 18.77 | 17.65 | **10.48** | 16.58 | 13.76 | **100.00** |

***Bolded** type denotes important prototype features

**TABLE** 2 Prominent Prototype Trip Origins and Destinations

| | ORIGINS | | DESTINATIONS | |
|---|---|---|---|---|
| Prototype | Community | % in Prototype | Prototype | % in Prototype |
| 1 | Near North Side | 22.62 | Near North Side | 24.22 |
| | Near West Side | 13.63 | Loop | 15.40 |
| | West Town | 9.638 | Near West Side | 14.15 |
| | Loop | 9.537 | West Town | 5.280 |
| | Lincoln Park | 5.878 | Lincoln Park | 5.012 |
| | Lake View | 5.169 | Lake View | 4.561 |
| 2 | Near North Side | 24.96 | Near North Side | 23.78 |
| | Near West Side | 12.84 | Near West Side | 13.16 |
| | Loop | 12.11 | West Town | 8.932 |
| | West Town | 7.502 | Lincoln Park | 8.162 |
| | Lincoln Park | 7.224 | Loop | 8.085 |
| | Lake View | 7.103 | Lake View | 7.664 |
| 3 | Near North Side | 16.77 | Loop | 18.21 |
| | Loop | 10.47 | Near North Side | 12.07 |
| | Lake View | 9.795 | Near West Side | 11.25 |
| | Near West Side | 8.263 | Lake View | 7.916 |
| | Lincoln Park | 7.144 | West Town | 4.967 |
| | West Town | 6.115 | Lincoln Park | 4.723 |
| 4 | **Midway*** | **13.80** | **O'Hare** | **16.17** |
| | **O'Hare** | **9.523** | **Midway** | **15.08** |
| | Near North Side | 7.606 | Near North Side | 9.966 |
| | Loop | 6.306 | Loop | 7.152 |
| | Near West Side | 5.624 | Near West Side | 6.974 |
| | Lake View | 4.607 | Lake View | 3.531 |
| 5 | **Loop** | **45.57** | **Loop** | **54.35** |
| | **Near North Side** | **32.15** | **Near North Side** | **21.88** |
| | Near West Side | 8.719 | Near West Side | 8.013 |
| | Lake View | 5.986 | Lake View | 5.982 |
| | West Town | 3.156 | West Town | 4.160 |
| | Lincoln Park | 2.688 | Lincoln Park | 3.112 |
| 6 | Near West Side | 13.81 | Near North Side | 14.90 |
| | Near North Side | 11.54 | Near West Side | 13.20 |
| | Loop | 10.60 | Loop | 13.18 |
| | West Town | 7.686 | West Town | 6.060 |
| | Lake View | 6.510 | Lake View | 6.058 |
| | Lincoln Park | 5.489 | Lincoln Park | 4.873 |

***Bold** type denotes important prototype features

## DISCUSSION

The K-prototypes analysis is geared at finding relationships in the ridesourcing data by grouping similar observations together. The merging of multiple datasets further enables the prototypes search to identify the main ridesourcing profiles with regards to trip attributes (e.g. travel time, fare, origin and destinations, being private or shared), and competing mobility services (transit and taxi) along with weather conditions. This discussion section focuses on how the results relate to current research and can inform future research directions. Four areas of investigation are highlighted, centering on weather impacts, competition with transit and taxi, ridesplitting patterns and spatial distribution of ridesourcing.

We find that while weather does not have a pervasive impact on ridesourcing across clusters, it does strongly determine the choices in P1 highlighted by its higher average windspeed, humidity, and rainfall in the last hour. The identification of prototype 1 gives evidence that weather can have a significant impact on TNC usage for as many as 24 % of trips. With the results from Frei et al. (2017) showing weather having an impact in a micro-transit focused choice experiments, the importance of including weather as an explanatory variable in future TNC analyses is illustrated. The use of weather in TNC analyses can further explain the interactions with ridesourcing and other modes. For example, weather was shown to impact active modes of transport, so including weather as an explanatory variable between the relationship of ridesourcing and active mobility can inform their demand in the future (Saneinejad et al., 2012). This is especially useful for understanding how TNCs might relate to bikeshare as adverse weather has been shown to decrease its demand and contribute to increased ridership of other modes (Gebhart and Noland, 2014). Brodeur and Nield (2018) find that ridesourcing demand increases during adverse weather conditions and compared the supply of TNC drivers to taxis. Their results illustrate the benefit of TNCs – in particular its dynamic pricing – over taxis as a tool to increase the supply of drivers and meet consumer demand. By utilizing more data like weather in studies that examine TNCs' relationship to other modes, it may reveal when TNCs are complementing or substituting those modes. A caveat of the current study is the focus on only one month of data. It is recommended for future work to explore longer temporal panels with a more complete sets of traditional and emerging modes to reveal the full extent of weather-related ridesourcing demand and substitution.

The importance of understanding the relationship TNCs have with other modes is further highlighted by P4 and P5. P4 shows that airport trips are a major source of demand for ridesourcing because it is more effective at serving it than current transit option for many users. As evidenced by the higher average transit travel times and low taxi demand, ridesourcing's advantage for these trips illustrates the need for careful planning with regards to airport amenities in the future. The presence of P4 highlights the importance for policymakers to determine the focus of airport improvements. Should O'Hare focus on increasing the capacity for ridesourcing pickups and drop-offs? Or should the city focus on improving transit connections to the airport? Mandle and Box (2017) find that TNCs have a major impact on airport services, and P4 shows more evidence for further research.

P5 illustrates the competitive nature beyond travel time of TNCs. Though table 1 shows that this is a smaller portion of the trips, representing only 5.1% of the data, this is still an interesting prototype because it emphasizes how TNCs offer several advantages that go beyond shorter travel times. With shorter transit travel times and the demand previously met by taxis, there must be factors such as comfort, safety, and convenience that must be considered in conjunction

with travel time. These other factors may determine how other modes can increase their competitiveness against TNCs.

Another major area of the literature is on the potential for TNCs to be a more efficient people mover than privately driven vehicles. The dynamic ridesharing literature examines the efficiency gains of ridesplitting over private modes (Xue et al., 2018, Alonso-Mora et al., 2017). Despite theoretical findings on the advantages of ridesplitting, there has been limited exploration of how this functions in real systems. A general result from this work worth mentioning is the low share of split rides despite a relatively high share of riders indicating that they are willing to share their ride. For the complete dataset, 26.7% of all trips were authorized to be shared but of these only 68.5% were actually shared. That implies that only 18.3% of the overall rides were truly pooled, likely reflecting a lack of matching travel itineraries that were close enough in space and time for the matching to occur. The percentage of authorized shared trips of all prototypes except for P6 is well below the 26.7% figure.

When compared to the other prototypes, P6 shows that pooled trip making can be seen as a separate profile of use. To further examine the patterns of ridesplitting, figure 3 shows the number of trips by separate trip-makers within a pooled trip for each prototype. P6 has a much higher share of shared trips including more than 3 riders. However, this prototype only constitutes 12.8% of the data. With such a small share of trips being shared, decision-makers that support TNCs should consider strategies that increase the number of pooled trips.

Lastly, we discuss the spatial distribution of travel. Notably, the majority of trips occur in or around the Chicago Loop or airports with standouts Near North Side and Near West Side where there are typically more residential units than the Loop and overall higher density compared to the rest of the city. **Table 2** confirms that the top 6 origins and destinations hardly differ across prototypes. The strong concentration of flows is further illustrated in **Figure 4** that shows the location of the top O-D pairs distinguished by bold borders. These areas tend to have higher influx of visitors, along with more leisure landmarks such as restaurants and night clubs. The residents of these community areas tend to have higher average incomes and possess higher educational attainments than the average Chicagoan. The ongoing debate in Chicago and cities around the US has focused on the lack of broader coverage, outside transit rich areas, of ridesourcing. Figure 4 highlights the lower share of rides occurring in and between historically underserved communities on the South and West sides of Chicago.

The existing research has identified that specific groups, that is, younger, better-educated, and more affluent individuals tend to use ridesourcing more (Rayle et al., 2014, Clewlow and Mishra, 2017). There is still a limited understanding of the socio-spatial equity of transformative mobility. A Seattle study found that UberX waiting times were (surprisingly) comparatively shorter in areas with lower income. Overall, the impact was larger in relation to population and employment density (Hughes and MacKenzie, 2016). Brown (2018) observes for Los Angeles that ridesourcing serves a broader set of neighborhoods and have lower cancellation rates and waiting times than traditional taxis. Future work should focus on gaining more understanding of how the overlapping disadvantages of transit coverage, user access needs (e.g. distance to work, amenities), digital divide and ethnic/income disadvantage impact and are impacted by ridesourcing diffusion and performance.

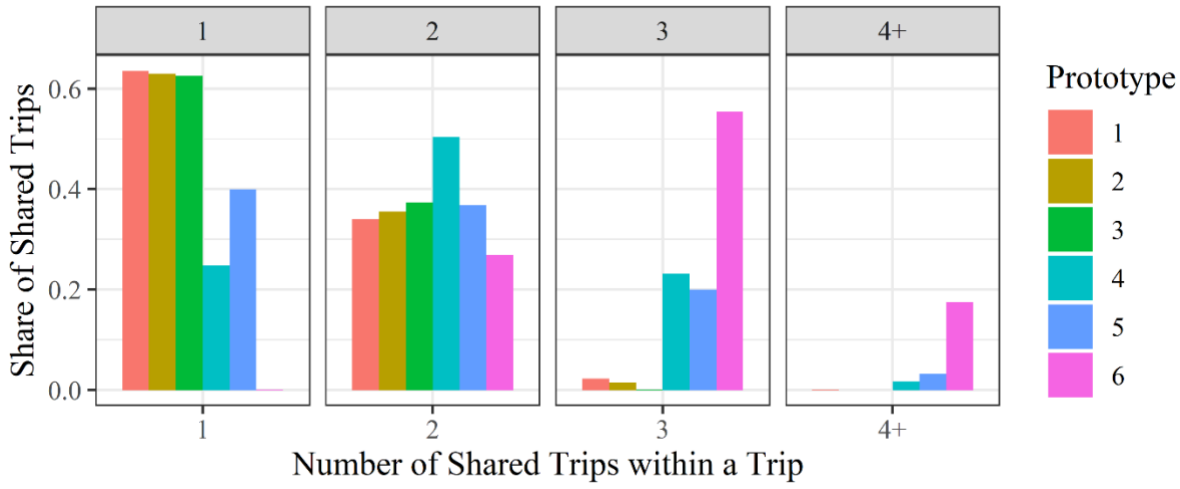


**FIGURE 3 Number of Travelers pooling a ride in Actual Shared Trips**

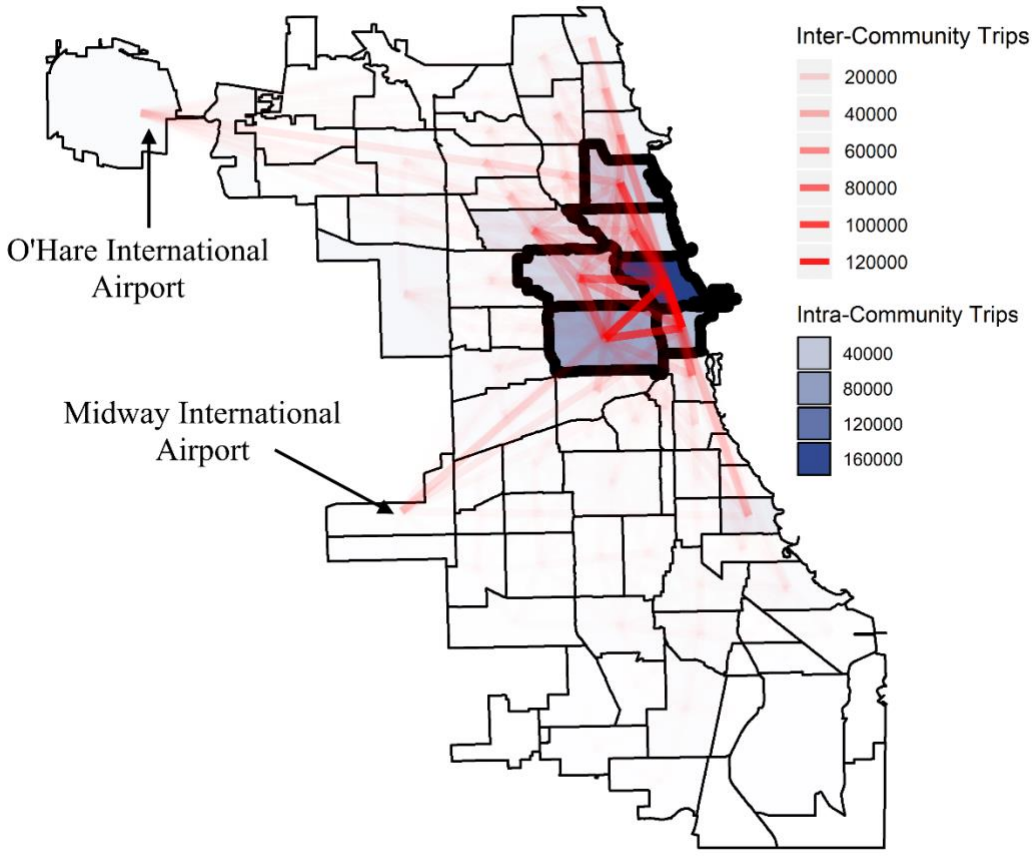


**FIGURE 4 Ridesourcing Flows in the City of Chicago with bolded boundaries of prominent Community Areas**

## CONCLUSION AND FUTURE WORK

This study examines a unique TNC dataset from Chicago, IL by utilizing an unsupervised learning algorithm that accepts categorical data. The use of K-prototypes showed that there were 6 distinct prototypes to describe underlying relationships in the data. Though demand and policy responses are not derived from this analysis, the goal of this study is to identify distinguishable features of TNC trips regarding service attributes, weather, transit, taxis, origins and destinations, and ridesplitting. The prototypes identified here shows that TNC trips can be described as a response to adverse weather conditions, evening trips, longer trips, trips to the airport, trips that could have happened just as fast as transit, or trips that are pooled.

The identification of these distinct trip types shows where future research is warranted. The discussion in this study focuses on how future research should consider factors such as weather and other external factors when estimating the demand for TNCs and other modes, airport-based mobility options in the future, understanding why TNCs have competitive advantages besides faster travel times, and why more trips are not shared. The last point made in the discussion is how most of the trips are completed in and surrounding the CBD of Chicago. The concentration of these trips in this area highlights the relationship between density and education and the role of TNCs in the broader picture of offering better mobility coverage (including a deeper discussion of where and for whom).

The main limitations of this study come from the constraints of limited data. Firstly, the weather data is collected at only one location. Considering the size of Chicago and the location of the station, the data may not be representative of local weather. Secondly, the TNC, taxi, and transit data are aggregated at the Census tract level. Because of this aggregation, the comparison of travel times between TNC and transit trips may not be as accurate. To increase the accuracy of these comparisons, more granular data is needed. Lastly, the TNC data only became available for trips from November 2018 and beyond. The data used in this project are only from November 2018, so more insights can possibly be gained by expanding the range of data.

## AUTHORS CONTRIBUTION

All authors contributed to all aspects of the study from model development, data processing, to analysis and interpretation of results, and manuscript preparation. All authors reviewed the results and approved the submission of the manuscript.

## ACKNOWLEDGEMENTS

Amanda Stathopoulos was supported in part by the US National Science Foundation Career Award No. 1847537.

## REFERENCES

ALEMI, F., CIRCELLA, G., HANDY, S. & MOKHTARIAN, P. 2018a. What influences travelers to use Uber? Exploring the factors affecting the adoption of on-demand ride services in California. *Travel Behaviour and Society,* 13**,** 88-104.

ALEMI, F., CIRCELLA, G., MOKHTARIAN, P. & HANDY, S. 2018b. Exploring the latent constructs behind the use of ridehailing in California. *Journal of Choice Modelling,* 29**,** 47-62.

1 ALONSO-MORA, J., SAMARANAYAKE, S., WALLAR, A., FRAZZOLI, E. & RUS, D. 2017.
2 On-demand high-capacity ride-sharing via dynamic trip-vehicle assignment. *Proceedings*
3 *of the National Academy of Sciences,* 114**,** 462-467.
4 BERGER, T., CHEN, C. & FREY, C. B. 2018. Drivers of disruption? Estimating the Uber effect.
5 *European Economic Review,* 110**,** 197-210.
6 BRODEUR, A. & NIELD, K. 2018. An empirical analysis of taxi, Lyft and Uber rides: Evidence
7 from weather shocks in NYC. *Journal of Economic Behavior & Organization,* 152**,** 1-16.
8 BROWN, A. E. 2018. *Ridehail Revolution: Ridehail Travel and Equity in Los Angeles*. PhD diss.,
9 University of California, Los Angeles.
10 CHEN, Y., HYLAND, M., WILBUR, M. P. & MAHMASSANI, H. S. 2018. Characterization of
11 Taxi Fleet Operational Networks and Vehicle Efficiency: Chicago Case Study.
12 *Transportation Research Record,* 2672**,** 127-138.
13 CIRCELLA, G., TIEDEMAN, K., HANDY, S., ALEMI, F. & MOKHTARIAN, P. 2016. What
14 Affects US Passenger Travel? Current Trends and Future Perspectives.
15 CITY OF CHICAGO 2019. Transportation Network Providers - Trips.
16 https://data.cityofchicago.org/Transportation/Transportation-Network-Providers-
17 Trips/m6dm-c72p.
18 CLEWLOW, R. R. & MISHRA, G. S. 2017. Disruptive transportation: The adoption, utilization,
19 and impacts of ride-hailing in the United States. *University of California, Davis, Institute*
20 *of Transportation Studies, Davis, CA, Research Report UCD-ITS-RR-17-07*.
21 CONTRERAS, S. D. & PAZ, A. 2018. The effects of ride-hailing companies on the taxicab
22 industry in Las Vegas, Nevada. *Transportation Research Part A: Policy and Practice,* 115**,**
23 63-70.
24 DIAS, F. F., LAVIERI, P. S., GARIKAPATI, V. M., ASTROZA, S., PENDYALA, R. M. &
25 BHAT, C. R. 2017. A behavioral choice model of the use of car-sharing and ride-sourcing
26 services. *Transportation,* 44**,** 1307-1323.
27 DONG, Y., WANG, S., LI, L. & ZHANG, Z. 2018. An empirical study on travel patterns of
28 internet based ride-sharing. *Transportation Research Part C: Emerging Technologies,* 86**,**
29 1-22.
30 FREI, C., HYLAND, M. & MAHMASSANI, H. S. 2017. Flexing service schedules: Assessing
31 the potential for demand-adaptive hybrid transit via a stated preference approach.
32 *Transportation Research Part C: Emerging Technologies,* 76**,** 71-89.
33 GEBHART, K. & NOLAND, R. B. 2014. The impact of weather conditions on bikeshare trips in
34 Washington, DC. *Transportation,* 41**,** 1205-1225.
35 GOOGLE. n.d. *Google Distance Matrix API* [Online]. Available:
36 https://developers.google.com/maps/ [Accessed].
37 HENAO, A. & MARSHALL, W. E. 2018. The impact of ride-hailing on vehicle miles traveled.
38 *Transportation*.
39 HUANG, Z. 1998. Extensions to the k-means algorithm for clustering large data sets with
40 categorical values. *Data mining and knowledge discovery,* 2**,** 283-304.
41 HUGHES, R. & MACKENZIE, D. 2016. Transportation network company wait times in Greater
42 Seattle, and relationship to socioeconomic indicators. *Journal of Transport Geography,* 56**,**
43 36-44.
44 JIANG, W. & ZHANG, L. 2018. The Impact of the Transportation Network Companies on the
45 Taxi Industry: Evidence from Beijing's GPS Taxi Trajectory Data. *IEEE Access,* 6**,** 12438-
46 12450.

KIM, K., BAEK, C. & LEE, J.-D. 2018. Creative destruction of the sharing economy in action: The case of Uber. *Transportation Research Part A: Policy and Practice,* 110**,** 118-127.

LEWIS, E. O. C. & MACKENZIE, D. 2017. UberHOP in Seattle: Who, Why, and How? *Transportation Research Record,* 2650**,** 101-111.

MADHULATHA, T. S. 2012. An overview on clustering methods. *arXiv preprint arXiv:1205.1117*.

MAHMOUDIFARD, S. M., KERMANSHAH, A., SHABANPOUR, R. & MOHAMMADIAN, A. 2017. Assessing public opinions on Uber and rdiesharing transportation systems: Exploratory analysis and results in a survey in Chicago. *2017 Annual Meeting of Transportation Research Board.* Washington D.C.

MANDLE, P. & BOX, S. 2017. *Transportation network companies: Challenges and opportunities for airport operators*.

NIE, Y. 2017. How can the taxi industry survive the tide of ridesourcing? Evidence from Shenzhen, China. *Transportation Research Part C: Emerging Technologies,* 79**,** 242-256.

OPEN WEATHER MAP. 2019. *Open Weather Map* [Online]. Available: https://openweathermap.org/ [Accessed].

R DEVELOPMENT CORE TEAM 2008. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing.

RAYLE, L., SHAHEEN, S., CHAN, N., DAI, D. & CERVERO, R. 2014. App-based, on-demand ride services: Comparing taxi and ridesourcing trips and user characteristics in san francisco university of california transportation center (uctc). *University of California, Berkeley, United States*.

SANEINEJAD, S., ROORDA, M. J. & KENNEDY, C. 2012. Modelling the impact of weather conditions on active transportation travel behaviour. *Transportation research part D: transport and environment,* 17**,** 129-137.

SCHWIETERMAN, J. & SMITH, C. S. 2018. Sharing the ride: A paired-trip analysis of UberPool and Chicago Transit Authority services in Chicago, Illinois. *Research in Transportation Economics,* 71**,** 9-16.

SHAHEEN, S. & CHAN, N. 2016. Mobility and the Sharing Economy: Potential to Facilitate the First- and Last-Mile Public Transit Connections. *Built Environment,* 42**,** 573-588.

SHAHEEN, S. & COHEN, A. 2018. Shared ride services in North America: definitions, impacts, and the future of pooling. *Transport Reviews***,** 1-16.

SZEPANNEK, G. & ASCHENBRUCK, R. 2019. clustMixType.

XUE, M., YU, B., DU, Y., WANG, B., TANG, B. & WEI, Y.-M. 2018. Possible Emission Reductions From Ride-Sourcing Travel in a Global Megacity: The Case of Beijing. *The Journal of Environment & Development,* 27**,** 156-185.

ZHEN, C. 2015. Impact of ride-sourcing services on travel habits and transportation planning.